Getting Feasible Variable Estimates From Infeasible Ones: MRF Local Polytope Study

Bogdan Savchynskyy
University of Heidelberg, Germany
bogdan.savchynskyy@iwr.uni-heidelberg.de

Stefan Schmidt
schmidt@math.uni-heidelberg.de

Abstract

This paper proposes a method for the construction of approximate feasible *primal solutions from infeasible ones for large-scale optimization problems possessing certain separability properties. Whereas the infeasible primal estimates can typically be produced from (sub-)gradients of the dual function, it is often not easy to project them to the primal feasible set, since the projection itself has a complexity comparable to the complexity of the initial problem. We propose an alternative efficient method to obtain feasibility and show that its properties influencing the convergence to the optimum are similar to the properties of the Euclidean projection. We apply our method to the local polytope relaxation of inference problems for Markov Random Fields and discuss its advantages over existing methods.*

1. Introduction

Convex relaxations of combinatorial problems appearing in computer vision, processing of medical data, or analysis of transport networks often contain millions of variables and hundreds of thousands of constraints. It is also quite common to employ their dual formulations to allow for more efficient optimization, which due to strong duality delivers also primal solutions. Indeed, approximate primal solutions can usually be reconstructed from (sub-)gradients of the dual objective. However, these are typically infeasible. Because of the problem size, only first order methods (based on the function and its (sub-)gradient evaluation only) can be applied. Since feasibility is not guaranteed up to the optimum, it is hardly attainable for such methods because of their slow convergence. The classical trick — (Euclidean) projection to the feasible set — can not be used efficiently because of the problem size.

A striking example of such a situation, which we explore in this paper, is the reconstruction of feasible primal estimates for local polytope relaxations of Markov random field (MRF) inference problems [32, 40, 37].

Motivation: Why Feasible Relaxed Primal Estimates Are Needed. It is often the case for convex relaxations of combinatorial problems that not a relaxed solution, but an integer approximation thereof is used in applications. Such integer primal estimates can be obtained from the dual ones due to the complementary slackness condition and using heuristic local search procedures [40, 16, 26]. However such integer estimates do not converge to the optimum of the relaxed problem in general.

In contrast, a sequence of *feasible* solution estimates of the *relaxed problem* converging to the optimum guarantees vanishing of the corresponding duality gap, and hence (i) determines a theoretically sound stopping condition [4]; (ii) provides a basis for the comparison of different optimization schemes for a given problem; (iii) allows for the construction of adaptive optimization schemes depending on the duality gap, for example adaptive step-size selection in subgradient-based schemes [18, 15] or adaptive smoothing selection procedures for non-smooth problems [29]. Another example is the tightening of relaxations with cutting-plane based approaches [35].

Related Work on MRF Inference. The two most important inference problems for MRF's are maximum a posteriori (MAP) inference and marginalization [37]. Both are intractable in general and thus both require some relaxation. The simplest convex relaxation for both is based on exchanging an underlying convex hull of the feasible set, the marginal polytope, by an approximation called the local polytope [37]. However, even with this approximation the problems remain non-trivial, though solvable, at least theoretically. A series of algorithmic schemes were proposed to this end for the local polytope relaxations of both MAP [18, 30, 26, 27, 33, 15, 29, 21, 20] and marginalization [36, 12, 10, 9]. It turns out that the corresponding dual problems have dramatically less variables and contain very simple constraints [40, 41], hence they can even be formulated as unconstrained problems as it is done in [30] and [15]. Therefore, most of the approaches address optimization of the dual objectives. A common difficulty for such approaches is the computation of a *feasible* relaxed primal

estimate from the current dual one. *Infeasible* estimates can typically be obtained from the subgradients of the dual function as shown in [18] or from the gradients of the smoothed dual as done in [13], [41], and [27].

Even some approaches working in the primal domain [10, 20, 33, 21] maintain infeasible primal estimates, whilst feasibility is guaranteed only in the limit.

Quite efficient primal schemes based on graph cuts proposed in [5] do not solve the problem in general and optimality guarantees provided by them are typically too weak. Hence we do discuss neither these here, nor the widespread message passing and belief propagation [16, 38] methods, which also do not guarantee the attainment of the optimum of the relaxed problem.

Forcing Feasibility of Primal Estimates. The literature on obtaining feasible primal solutions for MRF inference problems from infeasible ones is not very vast. Apart from the papers [27, 33, 29] describing special cases of our method in application to the MRF local polytope, we are aware of only three recent works [31, 42, 22] contributing to this topic.

The method proposed in [31] is formulated in the form of an algorithm able to determine whether a given solution accuracy ε is attained or not. To this end it restricts the set of possible primal candidate solutions and solves an auxiliary quadratic programming (QP) problem. However, this approach is unsuited to compute *the actually attained* ε directly and the auxiliary QP in the worst case grows linearly with the size of the initial linear programming problem. Hence obtaining a feasible primal solution becomes prohibitively slow as the size of the problem gets larger.

Another closely related method was proposed in [42]. It is, however, only suited to determine whether a given solution of the dual problem is an optimal one. This makes it non-practical, since the state-of-the-art methods achieve the exact solution of the considered problem only in the limit, after a potentially infinite number of iterations.

The very recent method proposed in [22] is simple yet efficient. However as we show in Section 2 (Theorem 2) our method applied on top of *any* other, including the one proposed in [22], delivers better primal estimates, except for the cases when the estimates of the other method coincide with ours.

Contribution. We propose an efficient and well-scalable method for constructing feasible points from infeasible ones for a certain class of separable convex problems. The method guarantees convergence of the constructed feasible point sequence to the optimum of the problem if only this convergence holds for their infeasible counterparts. In the case of the MRF local polytope our method coincides with the one proposed in [27]. We formulate our results in a general way, which allows to apply them to arbitrary convex optimization problems having a similar separable structure.

Content and Organization of the Paper. In Section 2 we describe a general formulation and analyze mathematical properties of the proposed method. We do this without relating it to inference in MRFs. This allows to keep the exposition simple and shows the generality of the method. Section 3 is devoted to the local polytope relaxations of the MAP and marginalization inference problems for MRF's and specifies how the feasible estimates can be constructed for them. In particular, in Section 3.4 we discuss different optimization schemes for the local polytope relaxation for which the primal estimates can be reconstructed from the dual ones. The last Sections 4 and 5 contain the experimental evaluation and conclusions, respectively.

Due to space constraints we refer to the technical report [28] for proofs of all theoretical results.

2. Optimizing Projection

Let us denote by $\Pi_C \colon \mathbb{R}^n \to C$ an Euclidean projection to a set $C \subset \mathbb{R}^n$. Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ be two subsets of Euclidean spaces and $C \subset X \times Y$ be a closed convex set. We will denote as C_X the set $\{x \in X \,|\exists y \in \mathcal{Y} \colon (x, y) \in C\}$, that is the projection of C to X.

The main definition of the paper introduces the notion of the *optimizing projection* in its general form. A possible simplification and the corresponding discussion follow the definition.

Definition 1. *Let $f \colon X \times Y \to \mathbb{R}$ be a continuous convex function of two variables. The mapping $\mathcal{P}_{f,C} \colon X \times Y \to C$ such that $\mathcal{P}_{f,C}(x, y) = (x', y')$ defined as*

$$x' = \Pi_{C_X}(x)\,, \tag{1}$$

$$y' = \arg\min_{y \colon (x',y) \in C} f(x', y)\,, \tag{2}$$

is called an optimizing projection *onto the set C w.r.t. the function f.*

This definition provides the way to get *the feasible* point $(x', y') \in C$ from an arbitrary infeasible one (x, y). Of course, getting just any feasible point is not a big issue in many cases. However, as we will see soon, the introduced optimizing projection possesses properties similar to the properties of a standard Euclidean projection, which makes it a useful tool in cases when its computation is easier than the one needed for the Euclidean projection. To this end both the partial projection (1) and the partial minimization (2) should be efficiently computable.

The role of projection (1) is to make x "feasible", i.e. to guarantee for x' that there is at least one $y \in \mathcal{Y}$ such that $(x', y) \in C$, which guarantees the definition to be well-defined. If this condition holds already for x, it is easy to see that $x' = x$ and hence computing (1) is trivial. We will call such x *feasible* w.r.t. C. Indeed, in (1) one can apply an arbitrary projection, since they all satisfy the mentioned

property. However, we provide our analysis for Euclidean projections only.

We will deal with objective functions, which fulfill the following definition:

Definition 2. *A function $f\colon X\times Y\to\mathbb{R}$ is called Lipschitz-continuous w.r.t. its first argument x, if there exists a finite constant $L_X(f)\geq 0$, such that $\forall y\in Y,\ x,x'\in X$,*

$$|f(x,y)-f(x',y)|\leq L_X(f)\|x-x'\| \tag{3}$$

holds. Similarly f is Lipschitz-continuous w.r.t.

- y if $|f(x,y)-f(x,y')|\leq L_Y(f)\|y-y'\|$ for all $x\in X,\ y,y'\in Y$ and some constant $L_Y(f)\geq 0$;
- $z=(x,y)$ if $|f(x,y)-f(x',y')|\leq L_{XY}(f)\|z-z'\|$ for all $z,z'\in X\times Y$ and some constant $L_{XY}(f)\geq 0$.

The following theorem specifies the main property of the optimizing projection, namely its continuity with respect to the optimal value of the function f.

Theorem 1. *Let f be convex and Lipschitz-continuous w.r.t. its arguments x and y and let f^* be the minimum of f on the set C. Then for all $z=(x,y)\in X\times Y$*

$$\begin{aligned}|f(\mathcal{P}_{f,C}(x,y))-f^*|&\leq|f(x,y)-f^*|\\&\quad+(L_X(f)+L_Y(f))\|z-\Pi_C(z)\|\end{aligned} \tag{4}$$

holds. If additionally x is feasible w.r.t. C the tighter inequality holds:

$$|f(\mathcal{P}_{f,C}(x,y))-f^*|\leq|f(x,y)-f^*|+L_Y(f)\|z-\Pi_C(z)\|\,. \tag{5}$$

Theorem 1 basically states that if the sequence $z^t=(x^t,y^t)\in X\times Y,\ t=1,\dots,\infty$, weakly converges to the optimum of f, then the same holds also for $\mathcal{P}_{f,C}(x^t,y^t)$. Moreover, the rate of convergence is preserved up to a multiplicative constant. Please note that $\mathcal{P}_{f,C}(x,y)$ actually does not depend on y, it is needed only for the convergence estimates (4) and (5), but not for the optimizing projection itself.

Remark 1. *Let us provide an analogous bound for the Euclidean projection to get an idea how good the estimate given by Theorem 1 is. Let $z^p=\Pi_C(z)$ denote the Euclidean projection of $z\in X\times Y$. Then*

$$\begin{aligned}|f(z^p)-f^*|&\leq|f(z^p)-f(z)|+|f(z)-f^*|\\&\leq|f(z)-f^*|+L_{XY}(f)\|z-z^p\|\,.\end{aligned} \tag{6}$$

We see that bounds (4) and (6) for the optimizing mapping and Euclidean projection differ only by a constant factor: in the optimizing mapping, the Lipschitz continuity of the objective f is considered w.r.t. to each variable x and y separately, whereas the Euclidean projection is based on the Lipschitz continuity w.r.t. the pair of variables (x,y).

The following technical lemma shows the difference between these two Lipschitz constants. Together with the next one it will be used in Section 3:

Lemma 1. *The linear function $f(x,y)=\langle a,x\rangle+\langle b,y\rangle$ is Lipschitz-continuous with Lipschitz constants $L_X(f)\leq\|a\|$, $L_Y(f)\leq\|b\|$ and $L_{XY}(f)\leq\sqrt{L_X(f)^2+L_Y(f)^2}$.*

Lemma 2. *The function $f(z)=\langle a,z\rangle+\sum_{i=1}^N z_i\log z_i$, where $\log$ denotes the natural logarithm, is Lipschitz-continuous at $[\varepsilon,1]^N\ni z,\ \varepsilon>0$, with Lipschitz-constant*

$$L_{XY}(f)\leq\|a\|+N|1+\log\varepsilon|\,. \tag{7}$$

An important property of the optimizing projection is its *optimality*. Contrary to the Euclidean projection it can deliver better estimates even when applied to already *feasible* point $(x,y)\in C$, which is stated by the following theorem.

Theorem 2 (Optimality of optimizing projection)**.** *Let $(x,y)\in C$ then $f(\mathcal{P}_{f,C}(x,y))\leq f(x,y)$ and the equality holds iff $y\in\arg\min_{y'\colon(x,y')\in C}f(x,y')$.*

Proof of this theorem follows directly from (2).

3. Inference Problems over Local Polytope and Corresponding Optimizing Projections

In this section we consider optimization problems related to inference in MRF's and construct corresponding optimizing projections. We switch from the general mathematical notation used in the previous sections to the one specific for the considered field, in particular we mostly follow the book of [37].

3.1. Primal Relaxed MAP Problem

Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be an undirected graph, where $\mathcal{V}$ is a finite set of nodes and $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$ is a set of edges. Let further $\mathcal{X}_v,\ v\in\mathcal{V}$, be finite *sets of labels*. The set $\mathcal{X}=\otimes_{v\in\mathcal{V}}\mathcal{X}_v$, where $\otimes$ denotes the Cartesian product, will be called *labeling set* and its elements $x\in\mathcal{X}$ are *labelings*. Thus each labeling is a collection $(x_v\colon v\in\mathcal{V})$ of labels. To shorten notation we will use x_{uv} for a pair of labels (x_u,x_v) and $\mathcal{X}_{uv}$ for $\mathcal{X}_u\times\mathcal{X}_v$. The collections of numbers $\theta_{v,x_v},\ v\in\mathcal{V},\ x_v\in\mathcal{X}_v$ and $\theta_{uv,x_{uv}},\ uv\in\mathcal{E},\ x_{uv}\in\mathcal{X}_{uv}$, will be called *unary* and *pairwise potentials*, respectively. The collection of all potentials will be denoted by θ.

Denoting $\mathbb{R}^{\sum_{v\in\mathcal{V}}|\mathcal{X}_v|+\sum_{uv\in\mathcal{E}}|\mathcal{X}_{uv}|}$ as $\mathbb{R}(\mathbb{M})$ and the corresponding non-negative cone $\mathbb{R}_+^{\sum_{v\in\mathcal{V}}|\mathcal{X}_v|+\sum_{uv\in\mathcal{E}}|\mathcal{X}_{uv}|}$ as $\mathbb{R}_+(\mathbb{M})$, one writes [32, 40] the local polytope (linear programming) relaxation of a MAP inference problem as

$$\begin{aligned}\min_{\mu\in\mathbb{R}_+(\mathbb{M})}&\sum_{v\in\mathcal{V}}\sum_{x_v\in\mathcal{X}_v}\theta_{v,x_v}\mu_{v,x_v}+\sum_{uv\in\mathcal{E}}\sum_{x_{uv}\in\mathcal{X}_{uv}}\theta_{uv,x_{uv}}\mu_{uv,x_{uv}}\\ \text{s.t. }&\begin{array}{l}\sum_{x_v\in\mathcal{X}_v}\mu_{v,x_v}=1,\ v\in\mathcal{V}\,,\\ \sum_{x_v\in\mathcal{X}_v}\mu_{uv,x_{uv}}=\mu_{u,x_u},\ x_u\in\mathcal{X}_u,\ uv\in\mathcal{E}\,,\\ \sum_{x_u\in\mathcal{X}_u}\mu_{uv,x_{uv}}=\mu_{v,x_v},\ x_v\in\mathcal{X}_v,\ uv\in\mathcal{E}\,.\end{array}\end{aligned} \tag{8}$$

The constraints in (8) form the *local polytope*, later on denoted as $\mathcal{L}$. Slightly abusing notation, we will briefly write problem (8) as $\min_{\mu\in\mathcal{L}} E(\mu) := \min_{\mu\in\mathcal{L}} \langle\theta,\mu\rangle$.

Optimizing Projection. We will denote as θ_w and μ_w, $w \in \mathcal{V}\cup\mathcal{E}$, the collections of θ_{w,x_w} and μ_{w,x_w}, $x_w \in \mathcal{X}_w$, respectively. Hence the vectors θ and μ become collections of θ_w and μ_w, $w \in \mathcal{V}\cup\mathcal{E}$. The n-dimensional simplex $\{x\in\mathbb{R}^n_+ \colon \sum_{i=1}^n x_i = 1\}$ will be denoted as $\Delta(n)$.

Problem (8) has a separable structure, that is for suitably selected matrices A_{uv} it can be written as

$$\begin{aligned}\min_{\mu\in\mathbb{R}(\mathbb{M})} &\sum_{v\in\mathcal{V}}\langle\theta_v,\mu_v\rangle + \sum_{uv\in\mathcal{E}}\langle\theta_{uv},\mu_{uv}\rangle \\ \text{s.t. }& \mu_v\in\Delta(|\mathcal{X}_v|), && v\in\mathcal{V}\,, \\ & A_{uv}\mu_{uv} = \mu_v,\ \mu_{uv}\geq 0, && uv\in\mathcal{E}\,.\end{aligned} \tag{9}$$

Note that under fixed μ_v, the optimization of (9) splits into small independent subproblems, one for each $uv\in\mathcal{E}$. We will use this fact to compute the optimizing projection onto the local polytope $\mathcal{L}$ as follows.

Let $\mu_\mathcal{V}$ and $\mu_\mathcal{E}$ be collections of primal variables corresponding to graph nodes and edges respectively, i.e. $\mu_\mathcal{V} = (\mu_v,\ v\in\mathcal{V})$, $\mu_\mathcal{E} = (\mu_{uv},\ uv\in\mathcal{E})$ and $\mu = (\mu_\mathcal{V},\mu_\mathcal{E})$. The corresponding subspaces will be denoted by $\mathbb{R}(\mathbb{M}_\mathcal{V})$ and $\mathbb{R}(\mathbb{M}_\mathcal{E})$. Then according to (9) and Definition 1, the optimizing projection $\mathcal{P}_{E,\mathcal{L}}\colon \mathbb{R}(\mathbb{M}_\mathcal{V})\times\mathbb{R}(\mathbb{M}_\mathcal{E})\to\mathcal{L}$ maps $(\mu_\mathcal{V},\mu_\mathcal{E})$ to $(\mu'_\mathcal{V},\mu'_\mathcal{E})$ defined as

$$\mu'_v = \Pi_{\Delta(|\mathcal{X}_v|)}(\mu_v),\ v\in\mathcal{V}\,, \tag{10}$$

$$\mu'_{uv} = \arg\min_{\substack{\mu_{uv}\geq 0 \\ \text{s.t. } A_{uv}\mu_{uv}=\mu'_v}} \langle\theta_{uv},\mu_{uv}\rangle\ ,\ uv\in\mathcal{E}\,. \tag{11}$$

Note that both (10) and (11) can be computed very efficiently. Projection to a simplex in (10) can be done *e.g.* by method described in [23]. The optimization problem in (11) constitutes a small-sized *transportation problem* well-studied in linear programming, see *e.g.* [2].

Let us apply Theorem 1 and Lemma 1 to the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ introduced in Definition 1. According to these, the convergence rate of a given sequence $\mu^t\in\mathbb{R}(\mathbb{M})$ in the worst case slows down by a factor $L_{\mathbb{M}_\mathcal{V}}(E) + L_{\mathbb{M}_\mathcal{E}}(E) \leq \|\theta_\mathcal{V}\| + \|\theta_\mathcal{E}\|$. This factor can be quite large, but since the optimum E^* grows together with the value $\|\theta_\mathcal{V}\|+\|\theta_\mathcal{E}\|$, its influence on the obtained *relative* accuracy is typically much less than the value itself.

Remark 2. *However, if θ contains "infinite" numbers, typically assigned to pairwise factors $\theta_\mathcal{E}$ to model "hard" constraints, both optimizing and Euclidean projections can be quite bad, which is demonstrated by the following simple example: $\mathcal{V}=\{v,u\}$, $\mathcal{E}=uv$, $\mathcal{X}_v=\mathcal{X}_u=\{0,1\}$, $\theta_{00}=\theta_{11}=\theta_{01}=0$, $\theta_{10}=\infty$. If now $\mu_{v,1}>\mu_{u,1}$, optimizing w.r.t. μ_{uv} leads to $\theta_{10}\cdot\mu_{vu,10} = \infty\cdot(\mu_{v,1}-\mu_{u,1})$, whose value can be arbitrary large, depending on the actual numerical value approximating ∞. And since neither the optimizing projection nor the Euclidean one take into account the actual values of pairwise factors when assigning values to $\mu_\mathcal{V}$, the relation $\mu_{v,1}>\mu_{u,1}$ is not controlled.*

We provide an additional numerical simulation related to infinite values of pairwise potentials in Section 4.

Remark 3 (Higher order models and relaxations)**.** *The generalization of the optimizing projection (10)-(11) for both higher order models, and higher order local polytopes introduced in [37, Sec. 8.5] is quite straightforward. The underlying idea remains the same: one has to fix a subset of variables such that the resulting optimization problem splits into a number of small ones.*

Remark 4 (Efficent representation of the relaxed primal solution)**.** *Note that since the pairwise primal variables $\mu_\mathcal{E}$ can be easily recomputed from unary ones $\mu_\mathcal{V}$, it is sufficient to store only the latter if one is not interested in specific values of pairwise variables $\mu_\mathcal{E}$. Because of possible degeneracy, there may exist more than a single vector $\mu_\mathcal{E}$ optimizing the energy E for given $\mu_\mathcal{V}$.*

3.2. Relaxed Dual MAP Problem

In this section we consider the Lagrange dual to the problem (8). Let us denote as $\mathcal{N}(v) = \{u\in\mathcal{V}\colon uv\in\mathcal{E}\}$ the set of neighboring nodes of a node $v\in\mathcal{V}$. We consider the dual variable $\nu\in\mathbb{R}(\mathbb{D})$ to consist of the following groups of coordinates: ν_v, $v\in\mathcal{V}$; ν_{uv}, $uv\in\mathcal{E}$; and $\nu_{v\to u,x_v}$, $v\in\mathcal{V}$, $u\in\mathcal{N}(v)$, $x_v\in\mathcal{X}_v$. In this notation the dual to (8) reads [32, 40]:

$$\max_{\nu\in\mathbb{R}(\mathbb{D})} \sum_{v\in\mathcal{V}}\nu_v + \sum_{uv\in\mathcal{E}}\nu_{uv} \tag{12}$$

$$\begin{aligned}\text{s.t. }& \theta_{v,x_v} - \textstyle\sum_{u\in\mathcal{N}(v)}\nu_{v\to u,x_v} \geq \nu_v\,,\ v\in\mathcal{V},\ x_v\in\mathcal{X}_v\,, \\ & \theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu_{v\to u,x_v} \geq \nu_{uv},\ uv\in\mathcal{E}\,, x_{uv}\in\mathcal{X}_{uv}\,.\end{aligned}$$

We will use the notation $\mathcal{U}(\nu) := \sum_{v\in\mathcal{V}}\nu_v + \sum_{uv\in\mathcal{E}}\nu_{uv}$ for the objective function of (12).

Optimizing Projection. The dual (12) possesses clear separability as well: after fixing all variables except ν_v, $v\in\mathcal{V}$, and ν_{uv}, $uv\in\mathcal{E}$, the optimization w.r.t. the latter splits into a series of small and straightforward minimizations over a small set of values

$$\nu_v = \min_{x_v\in\mathcal{X}_v}\theta_{v,x_v} - \textstyle\sum_{u\in\mathcal{N}(v)}\nu_{v\to u,x_v},\ v\in\mathcal{V}\,, \tag{13}$$

$$\nu_{uv} = \min_{x_{uv}\in\mathcal{X}_{uv}}\theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu_{v\to u,x_v},\ uv\in\mathcal{E}\,. \tag{14}$$

The formula (13) can be applied directly for each $v\in\mathcal{V}$, and (14) accordingly for each $uv\in\mathcal{E}$.

We denote by $\mathbb{D}$ the dual feasible set defined by constraints of (12). We split all dual variables into two groups.

The first one will contain "messages" $\nu_{\rightarrow} = (\nu_{v\to u},\ v \in \mathcal{V},\ u \in \mathcal{N}(v))$, that are variables, which reweight unary and pairwise potentials leading to improving the objective. The vector space containing all possible values of these variables will be denoted as $\mathbb{R}(\mathbb{D}_{\rightarrow})$. The second group will contain lower bounds on optimal reweighted unary and pairwise potentials $\nu_0 = (\nu_w,\ w \in \mathcal{V} \cup \mathcal{E})$. The total sum of their values constitutes the dual objective. All possible values of these variables will form the vector space $\mathbb{R}(\mathbb{D}_0)$. Hence the optimizing projection $\mathcal{P}_{\mathcal{U},\mathbb{D}} \colon \mathbb{R}(\mathbb{D}_{\rightarrow}) \times \mathbb{R}(\mathbb{D}_0) \to \mathbb{R}(\mathbb{D})$ maps $(\nu_{\rightarrow}, \nu_0)$ to $(\nu'_{\rightarrow}, \nu'_0)$ as

$$\nu'_{v\to u} = \nu_{v\to u},\ v \in \mathcal{V},\ u \in \mathcal{N}(v)\,, \tag{15}$$

$$\nu'_v = \min_{x_v \in \mathcal{X}_v} \theta_{v,x_v} - \sum\nolimits_{u \in \mathcal{N}(v)} \nu'_{v\to u,x_v},\ v \in \mathcal{V}\,, \tag{16}$$

$$\nu'_{uv} = \min_{x_{uv} \in \mathcal{X}_{uv}} \theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu'_{v\to u,x_v},\ uv \in \mathcal{E}\,. \tag{17}$$

Equation (15) corresponds to the projection (1), which has the form $\Pi_{\mathbb{R}(\mathbb{D}_{\rightarrow})}(\nu_{\rightarrow}) = \nu_{\rightarrow 0}$ and is thus trivial.

Applying Theorem 1 and Lemma 1 to the optimizing projection $\mathcal{P}_{\mathcal{U},\mathbb{D}}$ yields that the convergence of the projected ν^t slows down no more than by a factor $L_{\mathbb{D}_0} \leq |\sqrt{\mathcal{V}}| + |\sqrt{\mathcal{E}}|$ and does not depend on the potentials θ. However, since an optimal energy value grows often proportionally to $|\mathcal{V}| + |\mathcal{E}|$, the influence of the factor on the estimated related precision is typically insignificant.

3.3. Entropy-Smoothed Primal Problem

Let $H \colon \mathbb{R}^n_+ \to \mathbb{R}$ be *an entropy* function defined as $H(z) = -\sum_{i=1}^n z_i \log z_i$ and the dimensionality n will be defined by dimensionality of the input vector z. The problem

$$\min_{\mu \in \mathbb{R}_+(\mathbb{M})} \langle \theta, \mu \rangle - \sum_{w \in \mathcal{V} \cup \mathcal{E}} c_w H(\mu_w)$$
$$\text{s.t.} \begin{array}{l} \sum_{x_v \in \mathcal{X}_v} \mu_{v,x_v} = 1,\ v \in \mathcal{V}\,, \\ \sum_{x_v \in \mathcal{X}_v} \mu_{uv,x_{uv}} = \mu_{u,x_u},\ x_u \in \mathcal{X}_u,\ uv \in \mathcal{E}\,, \\ \sum_{x_u \in \mathcal{X}_u} \mu_{uv,x_{uv}} = \mu_{v,x_v},\ x_v \in \mathcal{X}_v,\ uv \in \mathcal{E}\,, \end{array} \tag{18}$$

is closely related to the primal relaxed one (8) and appears, *e.g.* when one applies the smoothing technique [24, 14, 27, 29, 10] to the problem or considers approximations for marginalization inference [37, 36, 12]. We refer to [11, 39, 10] for description of the sufficient conditions for convexity of (18). Assuming a precision $\varepsilon = 10^{-16}$ to be sufficient for practical needs, we equip (18) with an additional set of box constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$, where $|\mathbb{M}|$ is the dimensionality of the vector μ. This is done to obtain a finitely large Lipschitz constant according to Lemma 2.

Optimizing projection. Denoting the objective of (2) as $\hat{E}$ and the constraint equipped with the box-constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ as $\hat{\mathcal{L}}$ we define the corresponding optimizing projection $\mathcal{P}_{\hat{E},\hat{L}}(\mu)$ as

$$\mu'_v = \Pi_{\Delta(|\mathcal{X}_v|) \cap [\varepsilon,1]^{|\mathcal{X}_v|}}(\mu_v),\ v \in \mathcal{V}\,, \tag{19}$$

for $uv \in \mathcal{E}$:

$$\mu'_{uv} = \arg \min_{\substack{\mu_{uv} \in [\varepsilon,1]^{|\mathcal{X}_{uv}|} \\ \text{s.t. } A_{uv}\mu_{uv} = \mu'_v\,,}} \langle \theta_{uv} - c_{uv} \log(\mu_{uv}), \mu_{uv} \rangle \tag{20}$$

where $\log z$, $z \in \mathbb{R}^n$, is defined coordinate-wise. Applying Theorem 1 and Lemma 2 one obtains that the convergence rate of a given sequence $\mu^t \in \mathbb{R}(\mathbb{M})$ in the worst case slows down by a factor $\|\theta_{\mathcal{V}}\| + \|\theta_{\mathcal{E}}\| + \sum_{w \in \mathcal{V} \cup \mathcal{E}} |\mathcal{X}_w||1 + \log \varepsilon|$, where the last item describes a difference to the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ for the primal MAP-inference problem (8).

Remark 5. *Indeed, the additional constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ are needed only for the theoretical analysis of the projected estimate $\mathcal{P}_{\hat{E},\hat{L}}(\mu)$ to show that when the true marginals μ become close to* 0 *the optimizing projection (and Euclidean one indeed also) behaves worse.*

However there is no reason to force these constraints in practice: due to continuity of the entropy H the projected feasible estimates will converge to the optimum of the problem together with the non-projected unfeasible ones even without the box constraints. It is only the speed of convergence of the projected estimates, which will decrease logarithmically. Moreover, omitting the box constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ simplifies the computations (19) *and* (20)*. The first one corresponds then to projection to the simplex and the second one - to a small-sized* entropy minimization*, efficiently solvable by the Newton method after resorting to its corresponding* smooth *and* unconstrained *dual problem.*

Moreover, we suggest to threshold μ_v by setting μ_{v,x_v} to zero if it is less than the precision ε. It decreases the size of the subproblem (20) *and allows to avoid numerical problems.*

3.4. Application to Algorithmic Schemes

In previous sections we concentrated on the way to compute the optimizing projection assuming that a weakly converging (but infeasible) sequence is given. In this section we briefly discuss where these infeasible sequences come from.

Prox-Point Primal-Dual Algorithms (First-Order Primal-Dual, ADMM, ADLP). In the simplest case the (infeasible) optimizing sequences for the primal (8) and dual (12) problems are generated by an algorithm itself, as it is typical for primal-dual saddle-point formulation based algorithms. Some of these algorithms consider a slightly different dual formulation then (12) and maintain feasible dual variables [20, 21, 7], some do not [33]. However to the best of our knowledge none of these algorithms maintains feasibility of the primal estimates with respect to the problem (8). One can obtain the feasible estimates and respectively, the duality gap estimation, by applying

the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ defined by (10)-(11) and if needed $\mathcal{P}_{\mathcal{U},\mathbb{D}}$ defined by (15)-(17) respectively.

Subgradient Descent. Sub-gradient descent [1] was one of the first optimization algorithms with convergence guarantees, proposed in [30] and [17] for the dual problem defined by (12) and an equivalent dual based on the dual decomposition technique [17].

It is shown in [19] and later applied in [18] and [34, Sec.1.7.1] that time-averaged subgradients can be used to reconstruct the primal solution of the relaxed MAP-inference problem (8) and hence form the *infeasible* primal estimates, which can be turned to feasible ones with the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ defined by (10)-(11).

Methods Based on Smoothing/Methods for Marginalization Problem. There is a group of methods [27, 29, 12, 10, 26, 36, 13] addressing optimization of the entropy-smoothed primal problem (18) or its dual, which can be formulated as smooth and unconstrained one (see *e.g.* [27] for details). In the latter case gradient of the smooth dual function can be used to reconstruct infeasible primal estimates, as it is done in *e.g.* [27, 29]. Applying the optimizing projection $\mathcal{P}_{\hat{E},\hat{L}}$ defined by (19)-(20) provides feasible primal estimates converging to the optimum of the problem (18).

Remark 6. *If the final objective of the optimization is not the entropy-smoothed primal problem* (18)*, but the primal MAP-inference* (8)*, and the smoothing is used as an optimization tool to speed up or guarantee convergence [27, 29, 10, 13], one can obtain even better primal bounds for a lesser computational cost. Namely, the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ can be applied to approximate the optimal solution of the primal MAP-inference problem* (8)*. Denote $\hat{\mu}' = (\hat{\mu}'_{\mathcal{V}}, \hat{\mu}'_{\mathcal{E}}) = \mathcal{P}_{\hat{E},\hat{\mathcal{L}}}(\mu_{\mathcal{V}}, \mu_{\mathcal{E}})$ and $\mu' = (\mu'_{\mathcal{V}}, \mu'_{\mathcal{E}}) = \mathcal{P}_{E,\mathcal{L}}(\mu_{\mathcal{V}}, \mu_{\mathcal{E}})$.*

Ignoring the box-constraints according to recomendations of Remark 5, from the definitions (10) and (19) it follows that $\hat{\mu}'_{\mathcal{V}} = \mu'_{\mathcal{V}}$, and thus due to (11) and (20) $E(\mu') \leq E(\hat{\mu}')$. This means that the projection $\mathcal{P}_{E,\mathcal{L}}$ is preferable for approximating the minimum of E over $\mathcal{L}$ even in the case when the smoothed problem (18) *was optimized and not the original non-smooth* (8)*. As an additional benefit, one obtains faster convergence of the projection even from the worst-case analysis, due to a better estimate of the Lipschitz constant for the function E compared to the function $\hat{E}$, as provided by Lemmas 1 and 2.*

Non-smooth Coordinate Descent: TRWS, MPLP and others. We are not aware of methods for reconstructing primal solutions of the relaxed problem from dual estimates for non-smooth coordinate descent based schemes like TRW-S [16] and MPLP [8]. Indeed, these schemes do not solve the relaxed MAP problem in general, hence even if one would have such a method at hand, it would not guarantee convergence of the primal estimates to the optimum.

[1] We use the term *subgradient descent* also for maximization of concave functions

4. Experimental Analysis and Evaluation

The main goal of this section is to show how Theorem 1 works in practice. Hence we provide only two experiments to evaluate our method. Both concentrate on reconstructing of *feasible* primal estimates for the MAP inference algorithms considered in Section 3.4. In the first experiment we show how the projected primal MAP-solution converges to the optimum for three different algorithms. In the second one we show how the bound (4)-(5) allows for at least qualitative prediction of the objective value in the (feasible) projected point. We refer to [27, 33, 15, 29] for the experiments with an extended set of benchmark data.

For the experiments we used our own implementations of the First Order Primal Dual Algorithm (acronym **FPD**) [6] as described in [33], the adaptive diminishing smoothing algorithm **ADSAL** proposed in [29], the dual decomposition based subgradient ascent **SG** with an adaptive step-size rule according to [15, eq.17] and primal estimates based on averaged subgradients, and finally Nesterov's accelerated gradient ascent method **NEST** applied to smoothed dual decomposition based objective studied in [27]. All implementations are based on data structures of the OpenGM library [1].

The optimizing projection to the local polytope w.r.t. to the MRF energy (10)-(11) is computed using our implementation of a specialization of the simplex algorithm for transportation problems [2]. We adopted an elegant method by Bland [3], also discussed in [25], to avoid cycling.

Feasible Primal Bound Estimation. In the first series, we demonstrate that for all three groups of methods discussed in Section 3.4 our method efficiently provides feasible primal estimates for the MAP inference problem (8). To this end we generated a 256×256 grid model with 4 variable states ($|\mathcal{X}_v| = 4$) and potentials randomly distributed in the interval $[0, 1]$. We solved an LP relaxation of the MAP inference problem with **FPD** as a representative of methods dealing with infeasible primal estimates, subgradient method **SG** and **ADSAL** as the fastest representatives of smoothing-based algorithms. The corresponding plots are presented in Fig. 1. We note that in *all* experiments the time needed to compute the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ did not exceed the time needed to compute the subgradient/gradient of the respective dual function and required 0.01-0.02 s on a 3GHz machine. The generated dataset is not LP tight, hence the obtained relaxed primal solution has a significantly smaller energy than the integer one. In contrast to the cases where only non-relaxed integer primal estimates are computed, the primal and dual bounds of the relaxed problem converge to the same limit value. Due to the feasibility of both primal and dual estimates, the primal and dual

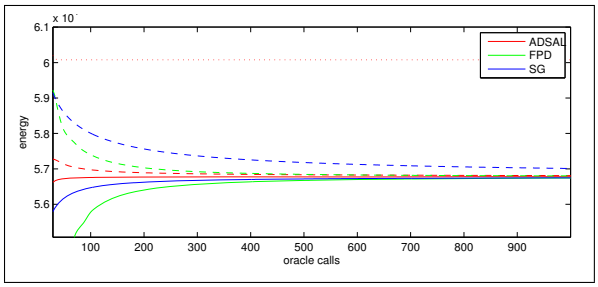

Figure 1. Convergence of the primal (dashed lines) and dual (solid lines) bounds to the same optimal limit value for **ADSAL** and **FPD** algorithms and **SG**. The obtained integer bound is plotted as a dotted line.

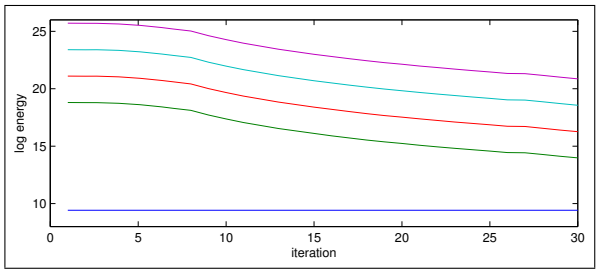

Figure 2. Convergence of the obtained primal feasible solution for four datasets which differ only by the values used as "infinity". The energy values are plotted in logarithmic scale. From bottom to top: optimal log-energy, primal bounds corresponding to infinity values equal to 10 000, 100 000, 1 000 000 and 10 000 000.

objective functions' values bound the optimal value of the relaxed problem from above and below, respectively.

Evaluation of Convergence Bound. The second experiment is devoted to the evaluation of the convergence bounds provided by Theorem 1. To this end, we generated four LP-tight grid-structured datasets with known optimal labeling. We refer to [33, pp. 95-96] for a description of the generation process. The resulting unary and pairwise potentials were distributed in the interval $[-10, 10]$. We picked up a random subset of edges not belonging to the optimal labeling and assigned them "infinite" values. We created four datasets with "infinities" equal to 10 000, 100 000, 1 000 000 and 10 000 000 and ran **NEST** for inference. According to Theorem 1 the energy E evaluated on projected feasible estimates $\mathcal{P}_{E,\mathcal{L}}(\mu^t_{\mathcal{V}}, \mu^t_{\mathcal{E}})$, $t = 1, \dots, \infty$, where the *infeasible* estimates μ^t were reconstructed from gradient of the dual function, can be represented as

$$E(\mathcal{P}_{E,\mathcal{L}}(\mu^t_{\mathcal{V}}, \mu^t_{\mathcal{E}})) = F(\mu^t) + L_Y(E)\|\mu^t - \Pi_{\mathcal{L}}\mu^t\| \quad (21)$$

for a suitably selected function F. Since **NEST** is a purely dual method and "infinite" pairwise potentials did not make any significant contribution to values and gradients of the (smoothed) dual objective, the infeasible primal estimates μ^t (with t denoting an iteration counter) were the same for all four different approximations of the infinity value. Since according to Lemma 1 the Lipschitz constant $L_Y(E)$ is asymptotically proportional to the values of the binary potentials $\theta_{\mathcal{E}}$ we plotted the values $\log E(\mathcal{P}_{E,\mathcal{L}}(\mu^t_{\mathcal{V}}, \mu^t_{\mathcal{E}}))$ as a function of t for all four datasets in Fig. 2. As predicted by Theorem 1 the corresponding energy values differ by approximately a factor of 10, as the "infinite" values do. Due to the logarithmic energy scale this difference corresponds to equal log-energy distances between the curves in Fig 2.

5. Conclusions

We presented an efficient and quite general optimizing projection method for computing feasible primal estimates for dual and primal-dual optimization schemes. The method provides convergence guarantees similar to the ones of the Euclidean projection, but contrary to it, it allows for efficient computations, when the feasible set and the objective function posses certain separability properties. As any optimization tool it has also certain limitations related to the Lipschitz continuity of the primal objective, however exactly the same limitations are characteristic also for the Euclidean projection. Hence they can not be considered as disadvantages of particularly this method, but rather as disadvantages of all projection methods in general and can be overcome only by constructing algorithms, which intrinsically maintain feasible primal estimates during iterations. The construction of such algorithms has to be addressed in future work.

Acknowledgement. This work has been supported by the German Research Foundation (DFG) within the program "Spatio-/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653.

References

[1] B. Andres, T. Beier, and J. H. Kappes. OpenGM: A C++ library for discrete graphical models. Technical report, arXiv:1206.0111, 2012.

[2] M. S. Bazaraa and J. J. Jarvis. *Linear Programming and Network Flows*. Wiley, 1977.

[3] R. Bland. New finite pivoting rules. *Discussion Paper 7612, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Heverlee, Belgium*, 1976.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, USA, 2004.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI, vol. 23, no. 11, pp. 1222-1239*, 23:1222–1239, 2001.

[6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2010.

[7] Q. Fu and H. W. A. Banerjee. Bethe-admm for tree decomposition based parallel map inference. In *UAI 2013*.

[8] A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*, 2007.

[9] T. Hazan, J. Peng, and A. Shashua. Tightening fractional covering upper bounds on the partition function for high-order region graphs. In *UAI*, 2012.

[10] T. Hazan and A. Shashua. Norm-product belief propagation: Primal-dual message-passing for approximate inference. *IEEE Trans. on Inf. Theory,*, 56(12):6294 –6316, Dec. 2010.

[11] T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413, 2004.

[12] J. Jancsary and G. Matz. Convergent decomposition solvers for tree-reweighted free energies. In *AISTATS*, 2011.

[13] J. K. Johnson, D. Malioutov, and A. S. Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *45th Ann. Allerton Conf. on Comm., Control and Comp.*, 2007.

[14] V. Jojic, S. Gould, and D. Koller. Accelerated dual decomposition for MAP inference. In *ICML*, pages 503–510, 2010.

[15] J. H. Kappes, B. Savchynskyy, and C. Schnörr. A bundle approach to efficient MAP-inference by lagrangian relaxation. In *CVPR 2012*, 2012. in press.

[16] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on PAMI*, 28(10):1568–1583, 2006.

[17] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[18] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. PAMI*, 33:531–552, March 2011.

[19] T. Larsson, M. Patriksson, and A.-B. Strömberg. Ergodic, primal convergence in dual subgradient schemes for convex programming. *Mathematical Programming*, 86:283–312, 1999.

[20] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. An augmented lagrangian approach to constrained MAP inference. In *ICML*, 2011.

[21] O. Meshi and A. Globerson. An alternating direction method for dual MAP LP relaxation. In *ECML/PKDD (2)*, pages 470–483, 2011.

[22] O. Meshi, A. Globerson, and T. S. Jaakkola. Convergence rate analysis of map coordinate minimization algorithms. In *NIPS*, pages 3023–3031, 2012.

[23] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of Rn. *J. Optim. Theory Appl.*, 50:195–200, July 1986.

[24] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, Ser. A(103):127–152, 2004.

[25] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Mineola, N.Y. : Dover Publications, 2nd edition, 1998.

[26] P. Ravikumar, A. Agarwal, and M. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *JMLR*, 11:1043–1080, 2010.

[27] B. Savchynskyy, J. Kappes, S. Schmidt, and C. Schnörr. A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In *CVPR 2011*, 2011.

[28] B. Savchynskyy and S. Schmidt. Getting feasible variable estimates from infeasible ones: MRF local polytope study. Technical report, arXiv:1210.4081, 2012.

[29] B. Savchynskyy, S. Schmidt, J. Kappes, and C. Schnörr. Efficient MRF energy minimization via adaptive diminishing smoothing. In *UAI 2012*, pages 746–755, 2012.

[30] M. Schlesinger and V. Giginyak. Solution to structural recognition (max,+)-problems by their equivalent transformations. in 2 Parts. *Control Systems and Computers*, (1-2), 2007.

[31] M. Schlesinger, E. Vodolazskiy, and N. Lopatka. Stop condition for subgradient minimization in dual relaxed (max,+) problem. In *EMMCVPR*, 2011.

[32] M. I. Schlesinger. Syntactic analysis of two-dimensional visual signals in the presence of noise. *Kibernetika*, (4):113–130, July-August 1976.

[33] S. Schmidt, B. Savchynskyy, J. Kappes, and C. Schnörr. Evaluation of a first-order primal-dual algorithm for MRF energy minimization. In *EMMCVPR 2011*, 2011.

[34] D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1, 2011.

[35] D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola. Tightening LP relaxations for MAP using message-passing. In *UAI*, pages 503–510, 2008.

[36] M. Wainwright, T. Jaakkola, and A. Willsky. A new class of upper bounds on the log partition function. *IEEE Trans. on Information Theory*, 51:2313–2335, 2005.

[37] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[38] Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, 47(2):736–744, 2001.

[39] Y. Weiss, C. Yanover, and T. Meltzer. Map estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence (UAI)*, 2007.

[40] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. on PAMI*, 29(7), July 2007.

[41] T. Werner. Revisiting the decomposition approach to inference in exponential families and graphical models. Technical report, CMP, Czech TU, 2009.

[42] T. Werner. How to compute primal solution from dual one in MAP inference in MRF? *Control Systems and Computers*, (2), March-April 2011.